Analysis of differentially expressed mRNAs and the prognosis of cholangiocarcinoma based on TCGA database

Kun Wang¹#, Yue Zhang²#, Xiaodan Yang³#, Tingsong Chen², Tao Han⁴⁵

¹Department of Hepatobiliary Surgery, Cancer Hospital of China Medical University/Liaoning Cancer Hospital, Shenyang, China; ²Department of Oncology II (Interventional Therapy), The Seventh People's Hospital of Shanghai University of Traditional Chinese Medicine, Shanghai, China; ³Department of Oncology, The General Hospital of Northern Theater Command, Shenyang, China; ⁴Department of Oncology, The First Affiliated Hospital of China Medical University, Shenyang, China; ⁵Department of Oncology, The Second Affiliated Hospital of Dalian Medical University, Dalian, China

Contributions: (I) Conception and design: All authors; (II) Administrative support: None; (III) Provision of study materials or patients: X Yang; (IV) Collection and assembly of data: K Wang, Y Zhang, X Yang; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

# These authors contributed equally to this work as co-first authors.

Correspondence to: Dr. Tao Han. Department of Oncology, The Second Affiliated Hospital of Dalian Medical University, Dalian, 116023, Liaoning, China. Email: than1984@sina.com. Dr. Ting-song Chen. Department of Oncology II (Interventional Therapy), The Seventh People's Hospital of Shanghai University of Traditional Chinese Medicine, Shanghai 200137, China. Email: chentingsong2243@163.com.

Background: The effective evaluation of cholangiocarcinoma (CCA) is challenging due to a lack of accurate screening tools. Consequently, there is an urgent need to screen out effective biomarkers. Bioinformatics analysis on a substantial amount of transcriptomic data to screen biomolecules allows for the verification of histological samples, and can provide a new method for CCA biomolecule screening in diagnosis and prognosis.

Methods: EdgeR model was used to analyze The Cancer Genome Atlas (TCGA)-extracted CCA data set, and to determine the differential expression of mRNAs. Based on this, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) were used to perform functional and pathway enrichment analysis. Subsequently, a protein interaction network was also established to identify the key differential node genes. Then, the previously determined differential genes were analyzed to establish a link between these genes and clinical prognosis. Finally, we used tissue samples to realize our results via IHC, Western blot and qRT-PCR.

Results: A total of 5,561 differential mRNAs were screened, including 3,473 upregulated genes and 2,088 downregulated genes. GO and KEGG enrichment analysis showed that the upregulated genes had significantly enriched cell adhesion, concentrated chromosomal motility, and microtubule motility. Downregulated genes were significantly enriched in heterologous metabolism and exosomes. Furthermore, we found upregulated genes were significantly enriched in the cancer pathways and cell cycle. Downregulated genes were enriched in the metabolic pathways and biosynthesis of antibiotics. Ten hub genes were screened out through the protein interaction network; among these, the AURKB and PLK1 genes were closely related to the clinical prognosis of patients. Results of the immunohistochemical staining, Western blot and qRT-PCR all showed that the expression of AURKB and PLK1 in cancer tissues was higher than that in the adjacent tissues, and this difference was statistically significant (P<0.05).

Conclusions: The upregulated genes were significantly enriched in the biological processes of cell division, cell cycle, and related cell components. AURKB and PLK1 play a key role in differentially expressed gene nodes. These genes are closely related to the prognosis of patients and can be used as potential diagnostic tools and prognostic biomarkers.
Introduction

Cholangiocarcinoma (CCA) is an abnormally aggressive malignant tumor originating in the bile duct epithelium and accounts for 10–20% of primary liver tumors. It is the second most common primary liver cancer, with a 5-year survival rate of only 5% (1). Recently, the incidence of CCA has increased each year. However, the incidence of CCA is occult, and when symptoms do appear, it is often impossible to operate (2–4). In most cases, the diagnosis of CCA in the early resectable stage is uncommon due to an insufficient understanding of the risk factors and the lack of accurate screening tools. CCA has a high incidence rate in many Asian countries, including China. At present, the carcinoembryonic antigen (CEA) and the carbohydrate antigens 19-9 (CA 19-9) and 125 (CA 125) are used clinically as serum markers for CCA; however, they have low sensitivity and specificity and are not adequate for providing an early detection method. For CA19-9, the most recent data resulting from a large meta-analysis that examined the distinction between CCA and healthy controls or patients with benign biliary disease, indicated pooled sensitivity and specificity of 72% and 84%, respectively (5). Similarly, the diagnostic sensitivity and specificity of CEA range from 42% to 85% and 70% to 89% respectively (6–8). In recent years, the research on non-coding RNAs as biomarkers have gradually increased. Circulating miR-21 is one of the most well-defined microRNAs and has the potential to be a biomarker for the diagnosis of CCA. Meanwhile, the study also showed that serum miR-21 levels were also positively correlated with tumor stage (TNM criteria) and poor survival. The level of serum miR-21 decreased after tumor resection, highlighting the value of circulating miR-21 as both a diagnostic tool and a putative prognostic biomarker (9). Non-coding RNAs with similar effects for CCA also include miR-26a, miR-106a and miR-150 (10,11). However, since most of the data on using circulating nucleic acids as biomarkers for diagnosis and prognosis of CCA come from conceptual experimental verification studies, further international assessment is urgently needed to confirm their potential clinical value.

Other promising biomarkers for circulating diagnosis and prognosis of CCA include cytokeratin 19 fragment (CYFRA 21-1), matrix metalloproteinase 7 (MMP-7) and osteopontin. Patients with intrahepatic CCA have increased CYFRA 21-1 compared with patients with benign biliary tract disease, and CYFRA 21-1 has a higher diagnostic value than CA19-9 and CEA (12). Also, the serum level of CYFRA 21-1 is related to the stage of the disease and is an independent predictor of the survival of patients without recurrence (12,13). Serum levels of MMP-7 in patients with CCA were also increased compared with benign biliary diseases, but its correlation with survival prognosis is currently unclear (14,15). Also, studies have shown that high levels of osteopontin in patients with CCA before and after surgery are associated with a decrease in overall survival (OS) after tumor resection (16). Overall, CCA currently lacks sensitive early diagnostic markers and effective prognostic biomolecules. Furthermore, the research on biomarkers is still scarce and needs further exploration. Our study uses a combination of bioinformatics and basic experiments, in which a protein interaction network is constructed to further narrow the scope of the key mRNAs. The purpose is to screen out key node molecules that regulate the poor prognosis of patients with CCA, and serve as targets for further mechanism research and therapeutic intervention.

Methods

Datasets and materials

We downloaded the mRNA expression data of CCA from The Cancer Genome Atlas (TCGA) database: 44 cases with genomic and corresponding clinical information were obtained from TCGA, including 35 cases of cancer tissue and 9 cases of normal tissue. Clinical information included age, gender, grade, stage, Eastern Cooperative Oncology Group (ECOG) score, family cancer history, OS time, and survival status. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Ethics Committee of the General Hospital of Northern Theater Command [No.: k(2017)12].
and informed consent was taken from all the patients. The total number of samples was 36, including 27 cancer tissues and 9 adjacent tissues. PLK1 antibody (ab17056) and Aurora B antibody (ab2254) were purchased from Abcam to check the expression of the corresponding molecules in the tissue by immunohistochemistry and Western blot.

**Differential gene analysis**

The R package, “edgeR”, was used to integrate the two sets of expression profile data to perform differential expression gene analysis. The genes with | log2 FC | >2 and false discovery rate (FDR) <0.05 were considered differentially expressed genes.

**Gene function enrichment analysis**

Through the R package, “ClusterProfiler” The Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) functions enriched by genes in the module of interest were obtained.

**Construction of protein interaction network and screening of Hub genes**

Cytoscape was used to characterize the interaction network between the differential genes and disease genes and to map the key protein interaction networks (highest confidence, 0.900). CytoHubba was used to calculate the top 10 key genes for the score.

**Survival analysis**

We divided the mRNAs into high and low expression groups by using the median score of the top 10 differential mRNAs combined with clinical data as the cutoff value; the log-rank test was used to test the differences in survival curves between the two groups and to screen for prognostic differential genes.

**Immunohistochemistry (IHC)**

IHC and results judgment were performed strictly in accordance with the immunochemistry operation process and standard specifications. PLK1 and AURKB staining showed brownish-yellow particles, which were cytoplasmic or nuclear and cytoplasmic. The immunohistochemically stained sections were observed under a microscope. An imaging device was used to take pictures of the fields of each group, and we randomly selected five classic fields after the picture. Pathologists scored them on the percentage of positively stained cells (A) and the intensity of staining of stained cells (B). The scoring procedure is described below.

(A) The number of positively stained cells was determined by first observing 5 high-power fields (+200) on each section and calculating the percentage of positive cells. Scoring was performed according to the percentage of positive cells as follows: <5%=0 points; 5–25%=1 point; 26–50%=2 points; 51–75%=3 points; and 76–100%=4 points.

(B) The scoring for positive coloring intensity was performed a follows: colorless =0 points; light yellow =1 point; brown yellow =2 points; and tan =3 points. The score of the two groups were added to obtain a total score for comparison.

**Western blot**

Extract the protein in the tissue, determine the protein concentration by quantitative method of BCA (bicinchoninic acid), and adjust each group of protein to the same concentration. After mixing with the loading buffer, the protein is denatured and transferred to the membrane. After blocking the non-specific antigen, the primary antibody was added and incubated overnight. The secondary antibody was diluted at a ratio of 1:2,000, and GAPDH was used as an internal control. The image was exposed, and was taken with Bio-RAD gel imaging system.

**qRT-PCR**

The RNA kit was used to extract the total RNA from the tissue samples, and the operation was carried out according to the instructions. Measure the total RNA concentration, synthesize cDNA according to the instructions of the reverse transcription synthesis kit, and then use the cDNA as a template to amplify the extracted RNA in the PCR instrument for reverse transcription. Real-time quantitative PCR was performed using SYBR Green dye method.

**Statistical methods**

All statistical analyses were performed using SPSS 19.0 software. Statistical analysis was conducted using a t-test or Bonferroni multiple comparisons test. A P value of less than
0.05 was considered statistically significant.

**Results**

**Identification of differential mRNA genes**

EdgeR analysis was used to analyze the different mRNA expressions between the normal tissues and tumor tissues (P<0.01); 5,561 differentially expressed genes were found, 3,473 of which were differentially upregulated mRNAs and 2,088 of which were differentially downregulated mRNAs (Figure 1A,B).

**Gene set enrichment analysis**

To further understand the effect of the screened differential genes on CCA, gene enrichment analysis was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) which includes GO and KEGG pathway enrichment analyses. Enrichment analysis was performed on up- and down-regulated genes. Through GO enrichment analysis of the upregulated genes, we observed many enriched gene sets. In terms of biological processes, they were enriched in cell adhesion and cell division. In terms of cellular components, they were significantly enriched in concentrated chromosomal motility and kinesin complexes. In terms of molecular function, they were mainly enriched in microtubule motility (Figure 2A). Through KEGG enrichment analysis of downregulated genes, we found that downregulated genes were mainly enriched in heterologous metabolism, exosomes, and oxidoreductase activity (Figure 2B).

We further analyzed the functional significance of differential mRNAs in the development of CCA through KEGG pathway analysis and discovered that upregulated genes were significantly enriched in the cancer pathway and
cell cycle (Figure 3A). Meanwhile, the downregulated genes were enriched in metabolic pathways and biosynthesis of antibiotics (Figure 3B).

**Protein interaction network**

A protein interaction network diagram of differentially expressed genes was constructed. The top 10 genes were selected as central genes according to degree and included CDK1, CDC20, CCNB1, CCNB2, CCNA2, BUB1, KNG1, PLK1, AURKB, and CDC18 (Table 1). Subsequently, the hub genes were submitted to STRING again to verify their interaction. The protein interaction network consisted of 10 nodes and showed the protein interactions between the closer hub genes (Figure 4). Then, the survival analysis was determined by a combination of score and clinical data, and the node genes that were closely related to survival prognosis were selected for verification.

**Survival analysis**

We made a matrix of differential mRNA expression and clinical survival time for the survival analysis. Then, we took the differential mRNAs with a P value <0.05 and obtained two differentially expressed genes: AURKB and PLK1 (Figure 5A,B,C,D,E,F,G,H,I,J).
Discussion

Although surgical resection is the primary treatment for CCA, most patients are found to already be in an advanced stage and are therefore unable to undergo an operation.

**Table 1** The top 10 mRNAs with scoring

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CDK1</td>
<td>87</td>
</tr>
<tr>
<td>2</td>
<td>CDC20</td>
<td>76</td>
</tr>
<tr>
<td>3</td>
<td>CCNB1</td>
<td>75</td>
</tr>
<tr>
<td>4</td>
<td>CCNB2</td>
<td>72</td>
</tr>
<tr>
<td>5</td>
<td>CCNA2</td>
<td>68</td>
</tr>
<tr>
<td>6</td>
<td>BUB1</td>
<td>65</td>
</tr>
<tr>
<td>7</td>
<td>KNG1</td>
<td>65</td>
</tr>
<tr>
<td>8</td>
<td>PLK1</td>
<td>63</td>
</tr>
<tr>
<td>9</td>
<td>AURKB</td>
<td>63</td>
</tr>
<tr>
<td>10</td>
<td>CDCA8</td>
<td>62</td>
</tr>
</tbody>
</table>

CDK1, cyclin-dependent kinases 1; CDC20, Cell division cyclin 20; CCNB1, CyclinB1; CCNB2, CyclinB2; CCNA2, CyclinA2; BUB1, budding uninhibited by benzimidazoles 1; KNG1, kininogen 1; PLK1, Polo-like kinase 1; AURKB, aurora kinase B; CDCA8, Cell division cycle associated 8.

**Figure 3** KEGG enrichment analysis showed that upregulated genes were significantly enriched in 31 pathways, including cancer pathway and cell cycle (A). Downregulated differentially expressed genes were significantly enriched in 37 pathways, including metabolic pathway (B). Count: number of genes related to the enriched GO or KEGG pathway. Term: enrichment pathway.

**Figure 4** The relationship network based on the top 10 hub genes.
Figure 5  The expression level of AURKB (A) and PLK1 (B) were both negatively associated with overall survival with a significant difference in TCGA (P<0.05). There is no significant difference in the survival time of high and low expression CDK1, CDC20, CCNB1, CCNB2, CCNA2, BUB1, KNG1, CDCA8 (C-J).
Thus, the early diagnosis of CCA is critical. Due to its occult clinical features and limited treatment measures, the prognosis of this malignant tumor is also poor. Therefore, it is critical to identify one or more biomarkers that could clarify the disease stage and prognosis. Carbohydrate antigen 19-9 (CA19-9) and cancer antigen 125 (CA125) have been widely used in most routine early detection tests for CCA. However, they have a broad range of sensitivity (50–90%) and specificity (54–98%) (17,18). Meanwhile, the increase in CA19-9 levels is also related to other metastatic diseases, and this is one of the reasons for its poor specificity as a serum biomarker for early diagnosis of CCA.

This study processed large amounts of data through bioinformatics methods to obtain more information about differentially expressed genes. In all, 5,561 differentially expressed genes were identified, 3,473 of which were...
differentially upregulated mRNAs, and 2,088 of which were differentially down-regulated mRNAs. GO enrichment analysis and KEGG enrichment analysis showed upregulated differential genes were significantly enriched in cell cycle and mitosis-related biological processes. We combined the survival data of 44 samples with the expression of the top ten key genes screened by CytoHubba, and performed a Kaplan–Meier survival analysis. Results from comparing the survival time between gene high and low expression groups, the results showed that the expression levels of Aurora kinase B (AURKB) and Polo-like kinase 1 (PLK1) were significantly correlated with OS. Results of the IHC also showed the expression levels of AURKB and PLK1 were significantly negatively correlated with OS. Gene expression also showed that the expression of AURKB and PLK1 in cancer tissues was higher than that in adjacent tissues (P<0.05).

Figure 8 Results of Western blot showed that the overall expression of PLK1 and AURKB in cancer tissues is higher than that in adjacent tissues. *, P<0.05.

Figure 9 At the level of gene, the qRT-PCR results also showed that the expression of PLK1 and AURKB in cancer tissues was higher than that in adjacent tissues (P<0.05).

Aurora kinases and Polo-like kinases are two crucial regulatory proteins in mitosis (19). Aurora kinases are a class of evolutionarily highly conserved serine/threonine kinases, in which the AURKB gene is the catalytic subunit of the chromosomal passenger complex (20). In recent years, research on AURKB in tumor-related fields has increased. Studies have shown that AURKB participates in the development of breast cancer, and its expression is related to the prognosis of breast cancer (21). In non-small cell lung cancer (NSCLC), overexpression of AURKB is associated with poor patient prognosis. It has been linked to the inhibition of p53-related pathways, which thereby promotes the proliferation of NSCLC cells, making it a possible mechanism of action (22). In colorectal cancer (CRC), AURKB has also been proven to be an essential oncogenic factor, which promotes drug resistance and progression of CRC (23,24). These studies collectively indicate that AURKB promotes canceration and is related to drug resistance, and that patients with high AURKB expression are more likely to have a poor prognosis (25).

Polo-like kinases are a family of serine/threonine kinases that participate in multiple biological processes, including mitosis, meiosis, and cytokinesis (26). Among them, PLK1 is a vital regulator of the eukaryotic cell cycle (27-29). PLK1 activity gradually increases during mitosis, reaches a peak at the G2/M phase, and enters the mitotic phase (30). PLK1 has different localization in different stages of mitosis, which determines its corresponding functions. PLK1 also acts as a mitotic switch to mediate repair after DNA damage, reducing genetic mutations caused by DNA damage (19). Similarly, PLK1 is highly expressed in most
human cancers, and its overexpression correlates with poor prognosis in cancer patients (31,32). The above findings are consistent with the results of functional enrichment analysis in this study, suggesting that AURKB and PLK1 molecules play key roles in regulating the proliferation of bile duct cancer cells.

Conclusions

Through bioinformatics analysis based on the TCGA database and verification of clinical tissue samples, we found that AURKB and PLK1 genes are differentially expressed in patients with CCA. These genes are closely related to the patients’ prognosis and could potentially be used as biomarkers for early diagnosis and prognosis. Furthermore, the bioinformatics analysis method can provide a reliable basis for early clinical diagnosis and the screening of prognostic biomarkers.

Acknowledgments

We thank our anonymous reviewers for their valuable comments on the manuscript, which contributed greatly to the improvement of the article.

Funding: This work was supported by grants from the Natural Fund Guidance Program of Liaoning Province (no. 2019-ZD-1072) and Key Projects of Liaoning Natural Science Foundation (no. 20180530019).

Footnote

Data Sharing Statement: Available at http://dx.doi.org/10.21037/tcr-20-812

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/tcr-20-812). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Ethics Committee of the General Hospital of Northern Theater Command [No.: k(2017)12] and informed consent was taken from all the patients.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References
